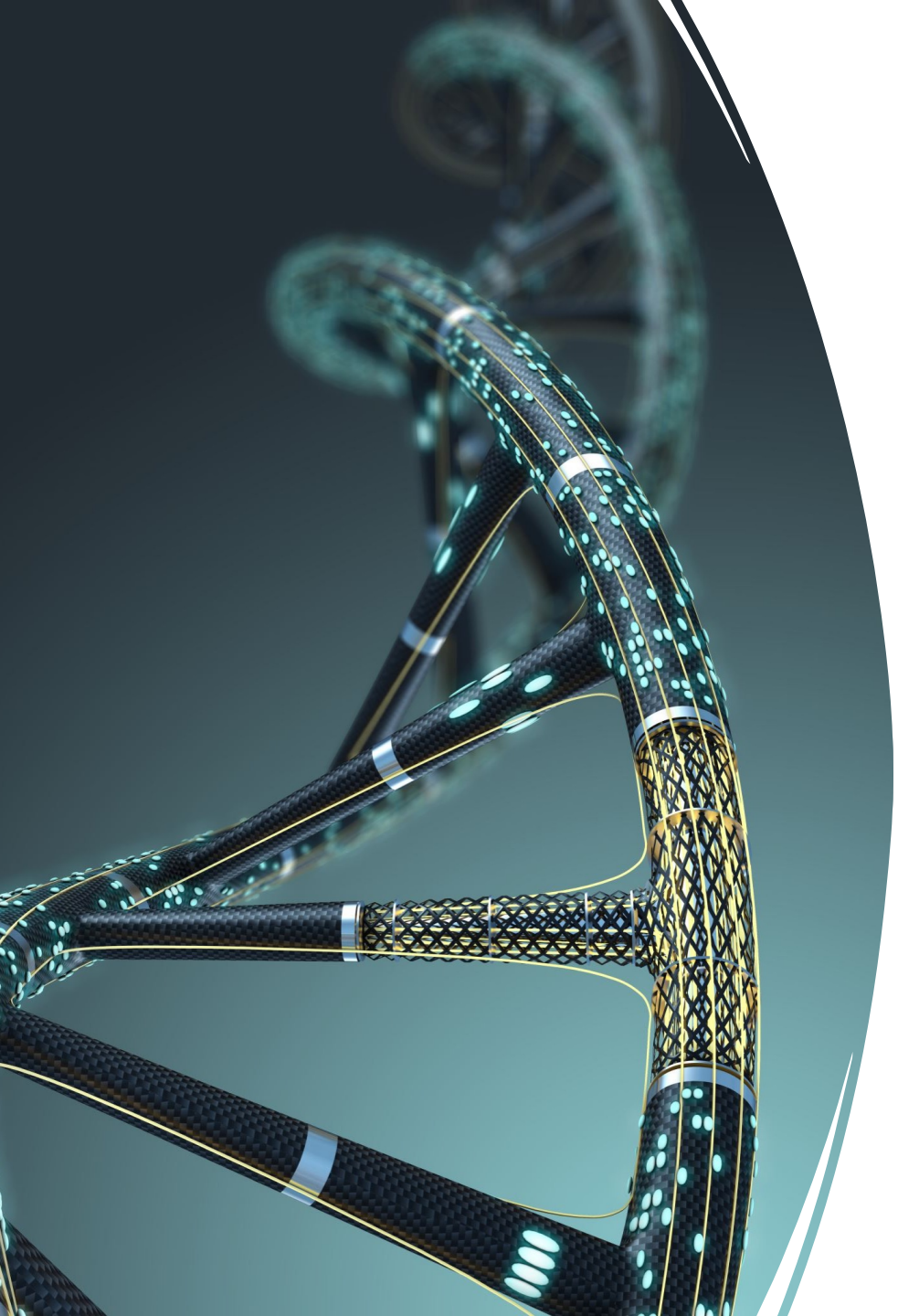


Predicting the Origin of a DNA Sequence Along with the Gene That It Encodes

By David Heffren and Sultan Daniels

Goal

Given a DNA sequence fragment, predict what organism this fragment came from.

Also, predict what gene this fragment encodes.

Why Study this Problem?

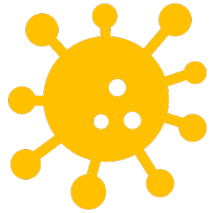
Metagenomics

The study of the genetic material present in a certain habitat.

As a large percentage of microbiological life is thought to still be undiscovered, DNA sequencing technology is an invaluable tool for better understanding our environment.

Evolutionary Biology

Accurate classification of DNA sequences can help scientists understand the genetic similarity between different species



Virology

As viruses are often too small to detect optically, DNA sequencing technology is used to detect, classify, and study viruses.

What is a Genome?

A genome is all the DNA of an organism.

- This includes genes and non-gene encoding DNA sequences

DNA Sequence

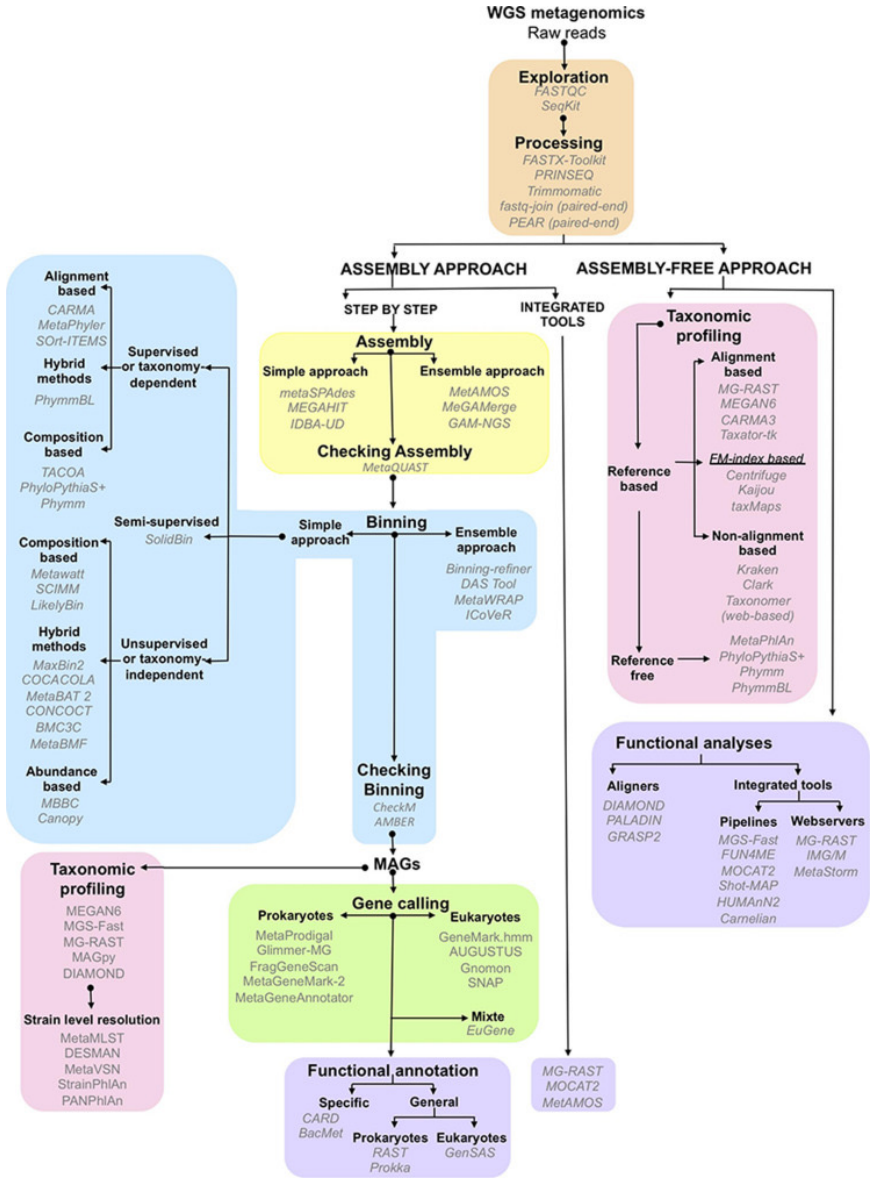
- A string of nucleotide bases

Genome Assembly

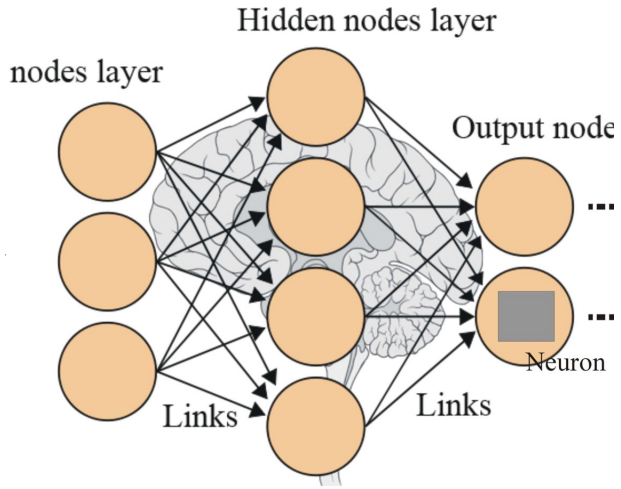
- Current DNA Sequencing technology does not have enough throughput to read out the sequence of an entire genome
- Consequently, genomes must be assembled using fragmented sequence

The Process of Conducting Metagenomic Studies

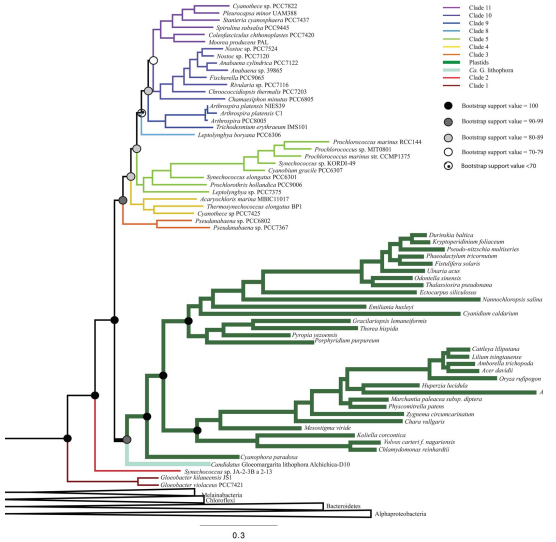
1. Sequencing the DNA
2. Assembling the Genome
3. Binning Similar Genetic Material
4. Annotating the Genome



Classification as a Binning Technique



Once the genome has been assembled it can be used as the training data for a neural network that will classify which organisms' specific sequences originated from.



The classification of these sequence fragments will allow for the partitioning of the environment's genetic material into different species.

Datasets

The assembled genomes of 4 different species of bear.

Each genome consists of about 2.3 GB of genetic data, partitioned into subsequences.

These assembled genome sequences are freely available on the website of the National Library of Medicine’s National Center for Biotechnology Information

As the different species of bear will most likely have significant genetic overlap, constraining our project to this scope will allow us to see how the neural network performs in a difficult case.

Structure of Our Data

- Our training data will consist of a list of subsequences of each bear genome with each subsequence given their class label.
- Due to memory limitations, these subsequences are chosen at random from each bear genome in order to get a representative sample of the entire genome.

Transforming Sequences into Numeric Form

In order to train our neural network, the training data must be in numeric form not nucleotide base characters.

In order to put these sequences into numeric form, we used k-mers which are analogous to k-shingles.

Each subsequence was represented as a vector where each component was the frequency of a corresponding k-mer occurring within the subsequence.

We chose to use k-mers in this way to preserve the information contained in the order of bases within subsequences.

K-mer Example

- DNA Sequence: GTAGAGCTGT
- Let k = 2
- K-mers: GT, TA, AG, GA, AG, GG, CT, TG, GT
- Unique k-mers: (GT, TA, AG, GA, GG, CT, TG).
- Frequency vector: (2, 1, 2, 1, 1, 1, 1).

Choice of k

As one can infer, the choice of k has a significant effect on the type of similarity that we are comparing in order to classify sequences.

In further iterations of our model, we will tune k as a hyperparameter.

Test Data

Currently, we are using subsequences of each genome that were not randomly selected to be in the training data as our test data.

We are still searching for a dataset of unassembled DNA sequence fragments of these bear species as more realistic test data.

```
Epoch 1/5
875/875 [==============================] - 8s 9ms/step - loss: 1.1482 - accuracy: 0.4880 - val_loss: 1.1365 - val_accuracy: 0.4760
Epoch 2/5
875/875 [==============================] - 7s 9ms/step - loss: 0.9733 - accuracy: 0.5886 - val_loss: 1.1403 - val_accuracy: 0.4807
Epoch 3/5
875/875 [==============================] - 7s 9ms/step - loss: 0.8850 - accuracy: 0.6345 - val_loss: 1.2106 - val_accuracy: 0.4733
Epoch 4/5
875/875 [==============================] - 8s 9ms/step - loss: 0.8265 - accuracy: 0.6609 - val_loss: 1.2424 - val_accuracy: 0.4555
Epoch 5/5
875/875 [==============================] - 8s 9ms/step - loss: 0.7883 - accuracy: 0.6814 - val_loss: 1.2842 - val_accuracy: 0.4552
250/250 [==============================] - 0s 2ms/step - loss: 1.2465 - accuracy: 0.4647
Model evaluation:  [1.2464550733566284, 0.4647499918937683]
```

Preliminary Results

- 46 % accuracy of classifying subsequences of bear genomes that were not selected to be in the training data.

Plan to Predict Gene Encoding

- For each of the 4 species of bear that we have data for, there is also an available dataset of their genomes with annotations.
- This means that gene encodings and other biological information is given along with the sequences that make up the genome.
- Our plan is to incorporate this information into our training data so that we can predict the gene encoding of a test DNA sequence.

References

Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol. 1998 Sep;180(18):4765-74. doi: 10.1128/JB.180.18.4765-4774.1998. Erratum in: J Bacteriol 1998 Dec;180(24):6793. PMID: 9733676; PMCID: PMC107498.

Castro CJ, Marine RL, Ramos E, Ng TFF. The effect of variant interference on de novo assembly for viral deep sequencing. BMC Genomics. 2020 Jun 22;21(1):421. doi: 10.1186/s12864-020-06801-w. PMID: 32571214; PMCID: PMC7306937.

Pérez-Cobas AE, Gomez-Valero L, Buchrieser C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. Microb Genom. 2020 Aug;6(8):mgen000409. doi: 10.1099/mgen.0.000409. Epub 2020 Jul 24. PMID: 32706331; PMCID: PMC7641418.

Pevsner, J. (2015). *Bioinformatics and functional genomics*. John Wiley & Sons.